Journal of Neuroscience Methods 208 (2012) 181–189

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

Detection and classification of subject-generated artifacts in EEG signals using autoregressive models

Vernon Lawhern[a,*], W. David Hairston[b], Kaleb McDowell[b], Marissa Westerfield[c], Kay Robbins[a]

[a] Department of Computer Science, University of Texas-San Antonio, San Antonio, TX 72849, USA
[b] Human Research and Engineering Directorate, US Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA
[c] Department of Neuroscience, University of California-San Diego, La Jolla, CA 92093, USA

HIGHLIGHTS

- We investigate the accurate detection and classification of subject-generated artifacts in continuous EEG recordings.
- Modeling of artifacts is performed using autoregressive (AR) modeling of artifact-contaminated EEG signals. Classification of EEG signals is performed using the support vector machine (SVM) classifier using the AR coefficients as features.
- Using the SVM classifier, we obtain accurate classification accuracy of a variety of artifacts (about 95%) across several subjects in our study.
- These results suggest that the AR coefficients can be used as features for classifying artifact-contaminated EEG segments.

ARTICLE INFO

Article history:
Received 13 March 2012
Received in revised form 15 May 2012
Accepted 16 May 2012

Keywords:
Autoregressive model
Artifacts
Electroencephalography
Support vector machines

ABSTRACT

We examine the problem of accurate detection and classification of artifacts in continuous EEG recordings. Manual identification of artifacts, by means of an expert or panel of experts, can be tedious, time-consuming and infeasible for large datasets. We use autoregressive (AR) models for feature extraction and characterization of EEG signals containing several kinds of subject-generated artifacts. AR model parameters are scale-invariant features that can be used to develop models of artifacts across a population. We use a support vector machine (SVM) classifier to discriminate among artifact conditions using the AR model parameters as features. Results indicate reliable classification among several different artifact conditions across subjects (approximately 94%). These results suggest that AR modeling can be a useful tool for discriminating among artifact signals both within and across individuals.

1. Introduction

In many EEG experimental paradigms, researchers attempt to limit the influence of outside effects on their experiments. For example, subjects are told not to perform any unnecessary movements or actions during the experiment, as these activities may confound the EEG activities of interest. After completing such a controlled experiment, researchers remove artifacts in EEG signals to obtain a "clean" signal that can be further analyzed. This process often requires manual identification of artifact-contaminated EEG, generally conducted by a panel of experts, which can be tedious and time-consuming, especially for large amounts of data. New applications of EEG are being performed in more complex and realistic environments, where controlling the effects of artifacts is not possible. An example of such an application is the detection of fatigue while driving (Lin et al., 2010), while EEG-based brain–computer interfaces (BCIs) are used to assist individuals with physical disabilities and to improve performance in healthy individuals (Lance et al., 2012). Off-line analyses that require extensive computation to remove artifacts are not feasible in these scenarios. Future applications of EEG to more realistic scenarios will require automated artifact detection methods that are robust to both inter-subject and intra-subject variations.

Autoregressive (AR) methods have been used in a number of studies to model EEG data by representing the signal at each channel as a linear combination of the signal at previous time points. In multivariate AR models, relationships between channels are also measured, providing useful information that can be used to calculate quantities such as ordinary, partial or directed coherence (Möller et al., 2001; Baccalá and Koichi, 2001) and the direct transfer function (DTF) (Franasczcuk et al., 1994). AR models provide a compact, computationally efficient representation of EEG signals. Furthermore, AR model parameters are invariant to scaling changes

* Corresponding author.
E-mail addresses: vlawhern@cs.utsa.edu, Vernon.Lawhern@utsa.edu (V. Lawhern), william.d.hairston4.civ@mail.mil (W.D. Hairston), kaleb.g.mcdowell.civ@mail.mil (K. McDowell), mwesterfield@ucsd.edu (M. Westerfield), krobbins@cs.utsa.edu (K. Robbins).

0165-0270
http://dx.doi.org/10.1016/j.jneumeth.2012.05.017

in the data that can arise from inter-subject variations, such as scalp and skull thickness. Due to these properties, AR modeling has been extensively used in EEG for different analyses such as feature extraction and classification tasks (Anderson et al., 1998), detection and classification of cardiac arrhythmias (Ge et al., 2002), and analysis of epilepsy data (Übeyli, 2010).

Previously, van de Velde et al. (1999) developed a method for detecting artifacts in EEG recordings. They combined several different measures of EEG activity including autoregressive (AR) parameter values, noise variance, and slope parameters to form features that are characteristic of artifact EEG. They categorized the artifacts as none, moderate, and severe. Using a panel of experts approach to tag data containing artifacts, they were able to accurately detect artifacts in EEG time series, with both high sensitivity (high true detection rate) and high specificity (low false detection rate). More recently, Chadwick et al. (2011) conducted a study on classifying eye and head movement artifacts in EEG using decision trees and Hidden Markov Models (HMMs). They used the mean, median, minimum, maximum, range and standard deviation of both the EEG signal and the first derivative of the signal as features for classifying among 13 different artifact groupings. Using the HMM, they report accuracies as high as 85% in some of their subjects.

In this paper, we propose a method for characterizing and classifying artifacts in EEG signals using coefficients obtained from an autoregressive model. An advantage to using AR modeling for artifacts is that the features obtained are scale-independent. The scale invariant feature of autoregressive models makes it possible to build a population feature set that characterizes the artifacts and allows decoding of artifacts on unobserved datasets despite the widely varying signal amplitudes and scales among trials and subjects. We also discriminate among different types of common EEG artifacts (eye blinks, saccades, muscle activity).

This paper builds on the progress made by previous work investigating the use of autoregressive models for discriminating artifacts within EEG data (e.g. van de Velde et al., 1999). In contrast to these earlier efforts, our work focuses on methods that are applicable across subjects. We omit the use of noise variance and slope parameters as these features are dependent on the scale of the data. We also apply multiple classifiers to discriminate among different types of artifacts. Finally, we build a population model of artifacts to decode artifact instances for subjects whose data was not part of the training data cohort. Our results indicate reliable discrimination among several different artifacts.

2. Materials and methods

2.1. EEG data collection and processing

Continuous EEG was recorded at 512 Hz using a 64-channel Biosemi ActiveTwo system (Biosemi, Amsterdam, Netherlands) and referenced to the average of the two mastoids. Four external channels were used to record eye movements by EOG. EOG activity was recorded to verify the instances of eye blinks and saccades in EEG, but was not used in subsequent analyses. The data were down-sampled to 256 Hz using a discrete wavelet transform with the Meyer wavelet family. This was done to reduce the computational burden as well to restrict the frequency range of analysis. The approximation coefficients of the down-sampled signal were then high-pass filtered at 1 Hz using an order 8 IIR Butterworth filter. We used EEGLAB (Delorme and Makeig, 2004) for processing and ERPLAB (Luck and Lopez-Calderon, 2011) for filtering the data.

2.2. Experimental setup

Seven participants performed a block of artifact-inducing facial and head movements, collected as part of a larger study. All provided consent prior to participating, and methods were approved as required by U.S. Army human use regulations (U.S. Department of the Army, 1990; U.S. Department of Defense Office, 1999). Before beginning, the experimenter reviewed the list of movements to ensure participants were familiar with the task. The exact details of each movement were not controlled by the experimenter, but rather left up to the participant to perform in their most natural manner. The seven movements included: clenching the jaw; moving the jaw vertically (like chewing gum); blinking both eyes (but without squinting); moving eyes leftward, then back to center; moving eyes upwards, then back to center; raising and lowering eyebrows; and rotating head side-to-side (as in looking leftward). All movements were performed sitting down in front of a PC screen. Each type of movement was performed in a separate run consisting of 20 repetitions. At the beginning of each run, instructions appeared on the screen reminding the participant of which movement should be made. For each run, a male voice initially counted down from 3 at a rate of every 2 s, followed by a tone every 2 s. Participants made the movement in time with the tone. The participants were told to make the movement for the first second of the 2 second period, and then to return to a relaxing state for the remaining 1 s.

A baseline dataset was also recorded for each participant, taken from the same larger study but during a different experimental run. The baseline dataset included three conditions: eyes open with the lights on, eyes open with the lights off, and eyes closed with the lights on. Participants were asked to look straight at a dot that was centered horizontally on top of the PC monitor during the "eyes open, lights on" and the "eyes open, lights off" conditions. We used data obtained from the "eyes open, lights on" condition as a baseline EEG for the presence of no artifact.

After data collection, epochs of length 500 ms, relative to the onset of each tone, were extracted from both the artifact dataset and the baseline dataset for each participant. Because participants were told to perform the movement in a natural fashion, variability was noted in the movement response latencies across subjects. For example; some participants waited for the audio tone to perform the movement, resulting in a reaction time delay of 300–400 ms, while other participants tried to predict the audio tone, resulting in some epochs not containing an artifact due to performing the movement too early. As a result, the epoch timing information was adjusted for each participant so that the time-course of the artifact was present in the epoch. Twenty non-overlapping 500 ms epochs were randomly extracted from the baseline dataset. We verified the absence of blink artifacts in these epochs by visual inspection. This process resulted in a total of 160 epochs, 20 for each of 8 conditions for each participant. An example plot of one 500 ms epoch for each artifact is shown in Fig. 1A for channel Cz. Here we see that most of the artifacts have somewhat identifiable time courses. For example, the Jaw Clench condition has a large amount of high frequency activity that can be easily differentiated from an eye blink. Differences in eye activity from the rest of the conditions are more noticeable for one of the frontal electrodes, as shown in Fig. 1B for channel Fp1. The eye blink has a more noticeable upward deflection in activity, while the eyebrow movement artifact has a much larger scale.

2.3. Statistical methods

2.3.1. Autoregressive models

The autoregressive framework assumes that the EEG signal can be modeled as a linear combination of the signals at the previous

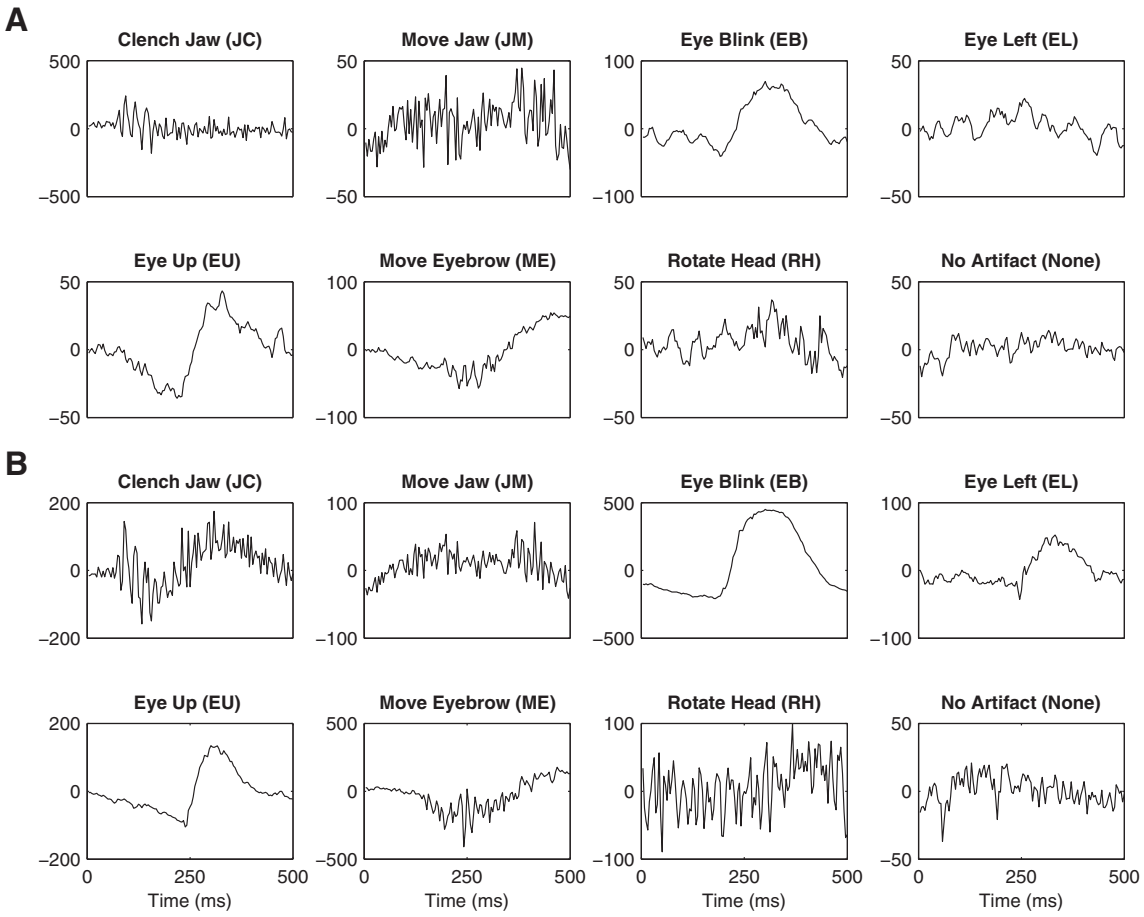

Fig. 1. (A) Plot of one epoch selected from each experimental condition, recorded at channel Cz, for one participant in the study. For the "no artifact" condition, one 500 ms epoch was randomly extracted from the baseline dataset. (B) Same as (A), but for channel Fp1. *Note*: The plots have different vertical scales.

time points. An autoregressive model of order p for a single channel can be written as:

$$y(t) = \sum_{i=1}^{p} \alpha_i y(t-i) + \epsilon_t \tag{1}$$

where p denotes the number of times points in the past that are used to model the current time point and ϵ_t denotes a zero-mean process with variance σ^2. The parameters of the AR model are the coefficients α_i, $i = 1, \ldots, p$ and the noise variance σ^2. Estimation of the AR parameters can be done using the Maximum Likelihood Estimator (MLE) for multiple linear regression models (Weisberg, 2005). Denote $\mathbf{Y}$ as the column vector of the time series data, and $\mathbf{X}$ as the matrix of covariates corresponding to the p previous time points. Then, the MLE can be written as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}) \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is a $p \times 1$ vector of parameter estimates that describe the characteristics of the signal.

We use the AR parameters as features for subsequent analyses due to their desirable properties. Namely, the AR parameters are invariant to the scale of the data because the scaling factor is cancelled out in the estimation of the parameters. If we scale the data by a constant c, then the MLE is:

$$\hat{\boldsymbol{\beta}} = (c\mathbf{X}^T c\mathbf{X})^{-1}(c\mathbf{X}^T c\mathbf{Y}) = (c^2\mathbf{X}^T\mathbf{X})^{-1}(c^2\mathbf{X}^T\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}) \tag{3}$$

For subsequent analyses, we use the AR model to model each EEG channel separately and concatenate the AR parameters to form a single feature vector that describes the time-series signal for a multi-channel EEG session.

2.3.2. Classification using support vector machines

We use a support vector machine (SVM) (Cortes and Vapnik, 1995) classifier to separate and classify the artifact signals in EEG. A study by Garrett et al. (2003) showed that the SVM performs optimally for a variety of classification scenarios. The features that we use in the SVM classifier are the single-channel AR coefficient estimates, which are concatenated together across channels to form a single vector. The goal of SVM is to construct the hyperplane (or hyperplanes) that optimally separates the data according to which class the data belongs to. In high dimensional spaces, kernel methods are used to keep the computational load reasonable (Garrett et al., 2003).

Here we present a summary on constructing the SVM model; details can be found in Chang and Lin (2011). For N total training vectors $\mathbf{x}_i$, $i, \ldots, N$, where the data belongs to one of two classes $y_i \in \{-1, 1\}$, SVM optimizes:

$$\begin{aligned} \underset{\omega,b,\epsilon}{\text{minimize}} \quad & \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{N}\epsilon_i \\ \text{subject to} \quad & y_i(\omega^T\phi(\mathbf{x}_i) + b) \geq 1 - \epsilon_i \\ & \epsilon_i \geq 0, \quad i = 1, \ldots, N. \end{aligned} \tag{4}$$

where $\phi(\mathbf{x}_i)$ is the mapping of $\mathbf{x}_i$ into a higher dimensional space, and C is the regularization term. Due to the computational

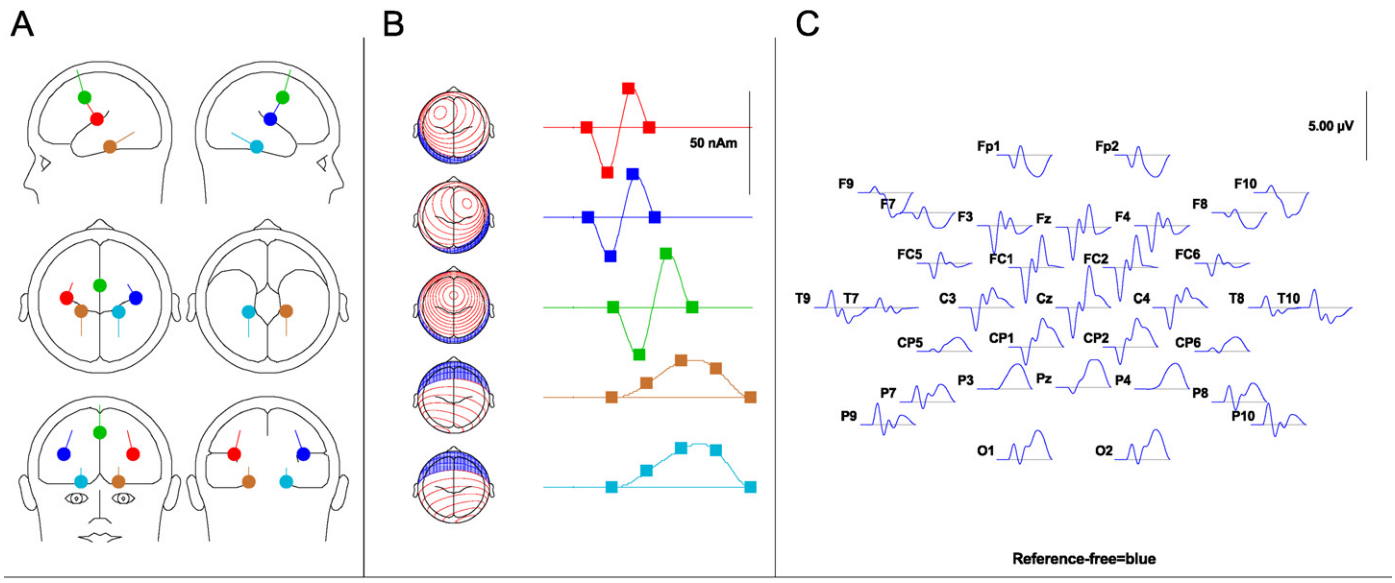

Fig. 2. The "n1p3b" dipole model from the DipoleSimulator program. (A) The locations of the 5 dipole sources in the head model. The brown–cyan and red–blue dipoles are symmetric across the head, while the green dipole is centered vertically on the head. (B) The detailed waveform characteristics for each of the dipole sources. (C) The estimated waveforms from the 33-channel cap based on the dipole model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

complexity of this optimization for high dimensional inputs, we perform a dual optimization (Boyd and Vandenberghe, 2004):

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ &\text{subjectto} \quad \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ &\qquad\qquad 0 \le \alpha^i \le C, \quad i = 1, \ldots, N. \end{aligned} \tag{5}$$

where $\mathbf{e} = [1, 1, \ldots, 1]^T$ is a vector of ones, $Q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, and $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel function. We use the radial basis function (RBF) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2). \tag{6}$$

where γ is the parameter of the RBF. Once Eq. (5) is optimized to produce $\boldsymbol{\alpha}$, we can find ω:

$$\omega = \sum_{i=1}^{N} y_i \alpha_i \phi(\mathbf{x}_i). \tag{7}$$

The decision rule is:

$$sgn(\omega^T(\phi(\mathbf{x})) + b) = sgn\left(\sum_{i=1}^{N} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \tag{8}$$

There are two parameters that need to be optimized: C, the regularization term and γ, the parameter in the RBF. We used a finely-partitioned grid search to find the optimal combination. We used the MATLAB toolbox LIBSVM (Chang and Lin, 2011) for building the SVM classifier.

For multi-class classification problems, LIBSVM uses a "one-against-one" approach: if M is the total number of classes, LIBSVM will build $M(M-1)/2$ classifiers, testing all the pairwise combinations of the data. A voting scheme is then implemented, where the label receiving the most votes is the classified label.

3. Results

3.1. Simulation experiments

We have shown above that the AR model parameters are invariant to scaling of the original data. To determine whether these parameters are invariant to subject variations in skull and scalp thickness, we used the BESA DipoleSimulator program (version 3.3.0.4, MEGIS Software GmbH, Gräfelfing, Germany) to build different head models. BESA allows the user to specify the dipole source locations and their waveforms as well as head model parameters such as scalp, skull, and cerebrospinal fluid (CSF) thickness. Once these parameters are specified, a specified EEG cap can be placed on the head model and EEG signals can be simulated according to the model specifications.

For the model simulations, we used the "n1p3b" dipole model included in BESA. As shown in Fig. 2, five distinct dipoles are modeled, each with different waveforms and locations (Fig. 2A and B). We used the 33-channel cap based on the international 10-10 system and simulated four different conditions where the scalp and skull thickness vary, reflecting individual differences in adult human subjects. Condition 1, the default condition, had scalp and skull thicknesses of 6 mm and 7 mm, respectively. Condition 2 had a decrease in skull thickness by 35% from baseline values; this value was chosen because a study on Korean adults observed variations in skull thickness as large as 35% along different regions of the skull (Hwang et al., 1997). Condition 3 had a reduction in scalp thickness of 15%, while skull thickness remained the same. Condition 4 had reductions in scalp and skull thicknesses of 15% and 35% of baseline, respectively. Thirty datasets were simulated at each condition.

We fit an AR(3) model to each channel in each condition and dataset separately. An order 3 model was chosen by using a Bayes Information Criterion (BIC) (Rissanen, 1989) model selection procedure; the average order was found to be 2.7 with a standard deviation of 1.3. An example of our simulation results are shown in Fig. 3. The first column is the histogram of all the AR(1) coefficient estimates from all the simulated EEG datasets across the four experimental conditions. The second and third columns are for the AR(2) and AR(3) coefficients, respectively. Here we see that the

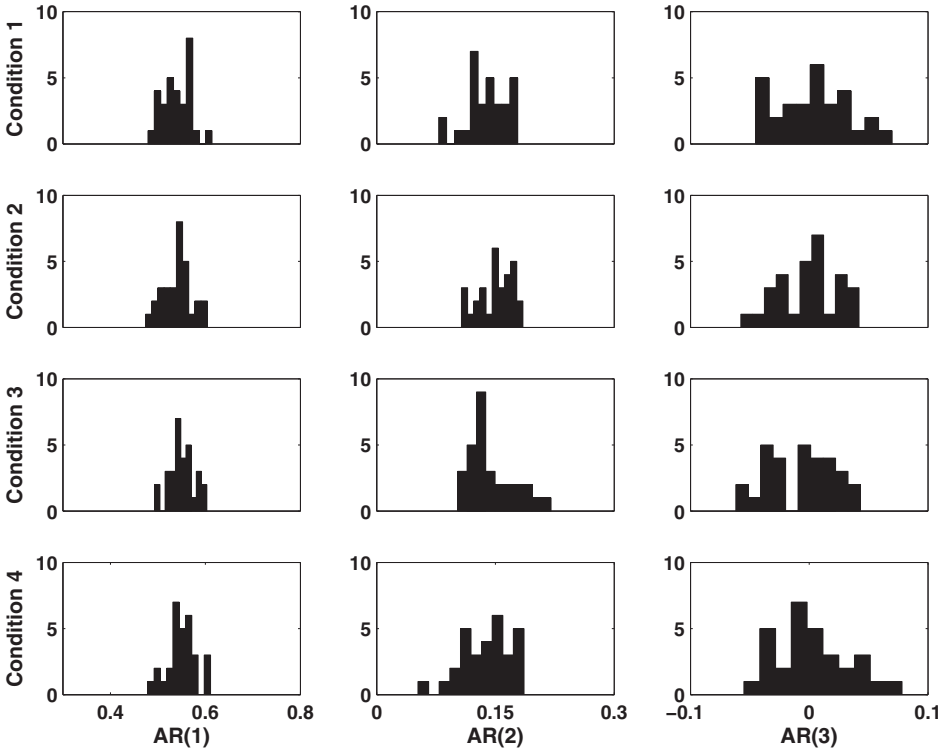

Fig. 3. Histograms of the AR parameters from channel F3 for 4 different selections of scalp and skull thickness. Condition 1: baseline, condition 2: 35% reduction in skull thickness, condition 3: 15% reduction in scalp thickness, condition 4: 35% reduction in skull thickness and 15% reduction in scalp thickness.

histograms for the AR coefficients are quite similar across experimental conditions.

We used a Kruskal–Wallis test to test the hypothesis that the parameters obtained from the AR models did not vary significantly across conditions within a specified channel. There are 33 tests, one Kruskal–Wallis test for each channel, testing the differences in coefficients across the four conditions for a specified model order. We used False Discovery Rate (FDR) analyses (Benjamini and Hochberg, 1995; Genovese et al., 2002) to adjust for the multiple testing problem. In FDR analysis, the experimenter gives a q-value, which is synonymous with the p-value for a single hypothesis test, to control the FDR for the family of tests being performed. Using FDR analysis, none of the 33 tests were significant for the AR(1), AR(2) and AR(3) parameters, respectively ($q = 0.05$), indicating no statistical difference among the four conditions for each of the three AR coefficients.

3.2. Classification of artifacts from real data

We concatenated the AR parameters from the individual channels to form a feature vector describing the data epoch. We divided the total dataset into two distinct parts of 60% and 40% for the purpose of training and testing, respectively. In the training data, we used 4-fold cross-validation to determine the best SVM model parameters and computed a cross validation (CV) accuracy. We then used the fitted model on the testing data, which is completely separate from the training process, to validate the model. The accuracy obtained from the test data is denoted as the testing accuracy (TA). This procedure was done 20 times on each subject, randomly partitioning the data at each iteration. Means and standard deviations across runs and across subjects were calculated for CV, TA, and individual artifact classification accuracies.

We fitted AR models to each channel individually for each data epoch. High classification accuracy was observed when fitting relatively small model orders (ranging from 1 to 4). We used an AR(2) model for all the channels as it gave high classification accuracy at low computational cost (see Fig. 4). Classification performance for higher orders was not significantly different than the performance of the AR(2) models.

Fig. 5 shows two-dimensional projections of the 128-element AR features vectors from three participants (A–C). Each point in this plot represents the AR coefficient estimates corresponding to one channel and one epoch. As seen in Fig. 5A, there is noticeable structure to the AR coefficients corresponding to different artifact

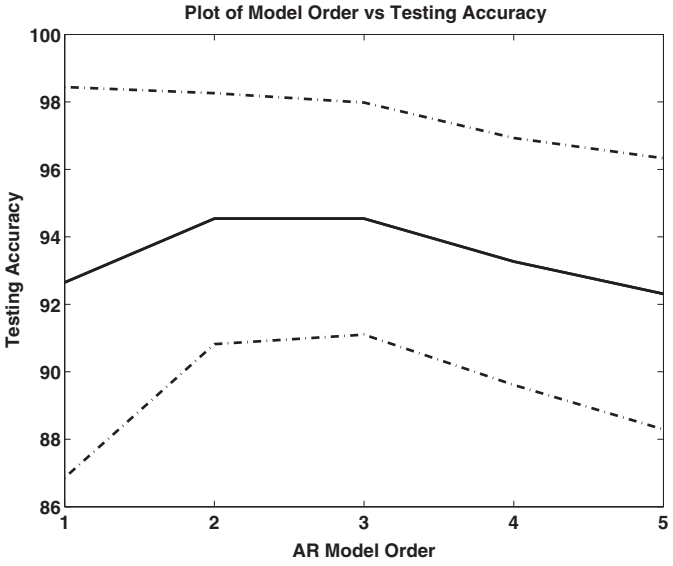

Fig. 4. Plot of the testing accuracy (TA) vs. the AR model order. The solid line denotes the testing accuracy, while the dashed lines denote 1 standard deviation of the testing accuracy.

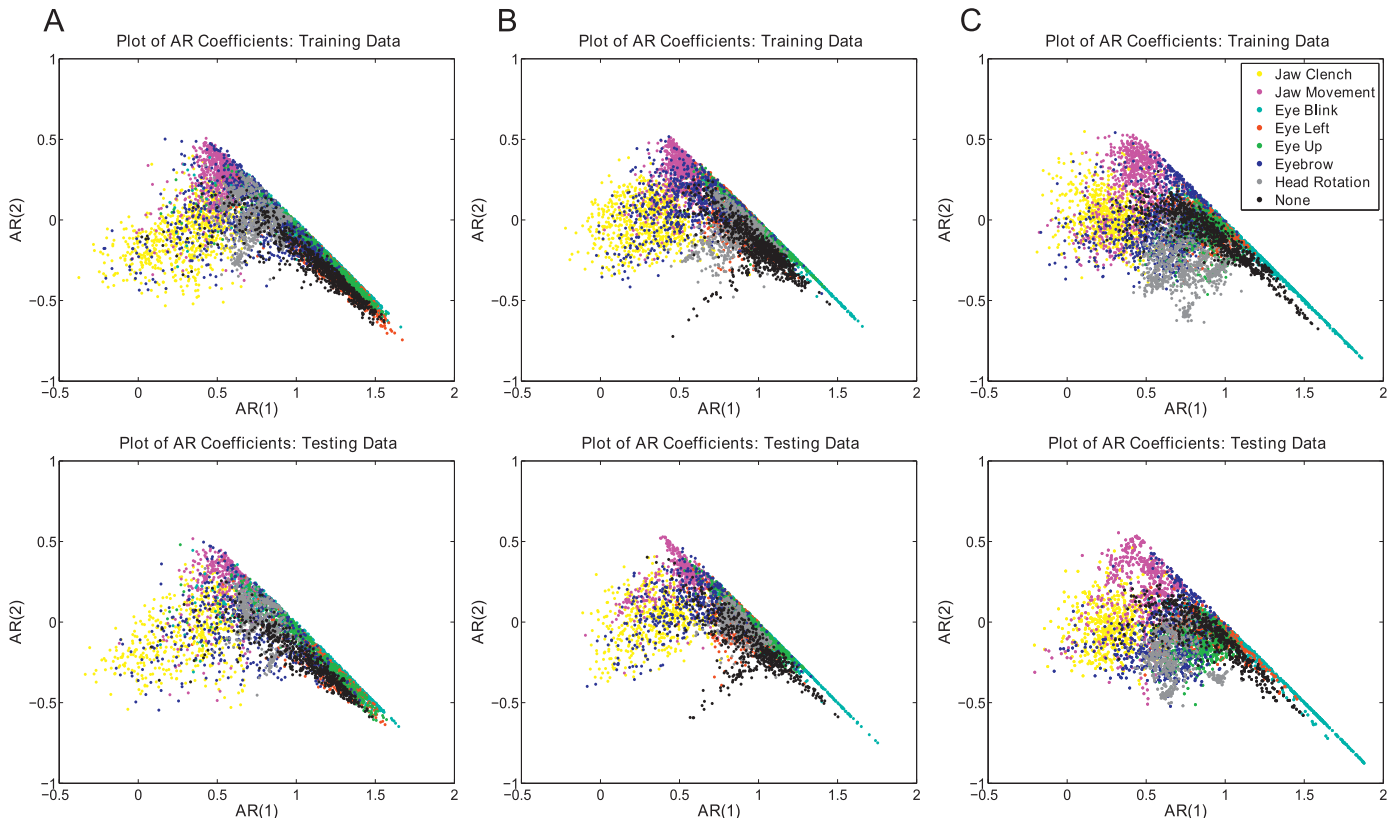

Fig. 5. Plot of the AR(1) vs. AR(2) coefficients for three participants in the study (**A**, **B** and **C**). Each point in these plots corresponds to the coefficients for all channels for all epochs. The first row of figures denotes the parameters from the training set, in which 60% of the epochs are randomly selected from the total dataset. The second row of figures denotes the parameters from the testing set, the remaining 40%.

conditions. For example, both the Jaw Clench and Jaw Movement conditions are grouped together, while the eye artifacts (blinks and vertical/horizontal saccades) are grouped and are separated from the jaw artifact group. This is expected, as the jaw artifacts exhibit high frequency muscle activity that is noticeably different from eye movement. The vertical eyebrow movements contain both muscle activity from the forehead as well as eye activity reminiscent of saccades. The vertical eyebrow coefficients are somewhere between the Jaw Clench/Movement group and the eye activity group. The head rotation condition usually has subjects moving their eyes to adjust for the visual orientation of the head, and these features are fairly close to the eye saccade conditions. The epochs with no artifact are grouped together in the middle of the 2-dimensional space. From the figure we also see that the features are consistent across the two datasets (training in the first row and testing in the second row) with regions from the training set being accurately reflected in the testing set. Fig. 5B and C shows the plots of the AR coefficient features for two other participants in the study. All three participants appear to have similar feature spaces for all the artifact conditions as well as the no artifact condition.

Fig. 6 shows the results of several different artifact classification analyses that we performed. Fig. 6A shows the performance for the SVM classifier within subjects, but averaged across all subjects. The rows of the matrix denote the original label, while the columns denote the classified label as determined by the SVM classifier. Diagonal entries denote correct classification, while off-diagonal entries denote misclassification. From Fig. 6A, we see that the entries are mostly diagonal, indicating that the within-subject classification rate of artifacts is high (numerical results are shown in Table 1). A few epochs from the Jaw Movement condition are incorrectly classified as Jaw Clench; this is reasonable in the sense that both conditions contain muscle artifact. Overall, most of the artifacts are classified correctly.

We tested for the significance of the observed classification probabilities by using a bootstrap permutation test, where at each bootstrap sample, we randomly permutated the labels (here, the artifact types) while keeping the features fixed. After several iterations of this procedure, we created an approximate 95% confidence interval by taking the mean and standard deviation of the bootstrap samples. This permutation effectively eliminates any relationship that the features may have had in distinguishing between the artifact types, thereby testing the hypothesis that the features are independent of the artifact type. The bootstrap permutation test results are shown in Table 1. The classification probabilities with the permutation test are not significantly different from a random classification (12.5%, or 1/8) and are significantly different than either the average or pooled classification performance.

Fig. 6B shows the results of a pooled data analysis, where all the data from the subjects were combined together to form a new dataset. We used the same analysis structure for this combined dataset (60% training, 40% testing, 20 repetitions, then average over the repetitions). We see a similar structure as in Fig. 6A; the highly diagonal responses indicate high classification probability for the pooled dataset. A few responses were incorrectly classified; notably, some of the head rotation responses were classified as leftward eye movements. This was expected since the head rotation condition had subjects rotating their head 90° to the left then back to center; the eye activity corresponding to the head rotation may be classified as an eye movement. Similarly, some blink epochs were classified as upward eye movements. These two activities have similar EEG patterns that could potentially lead to misclassification. Over 95% of the epochs from the no artifact condition

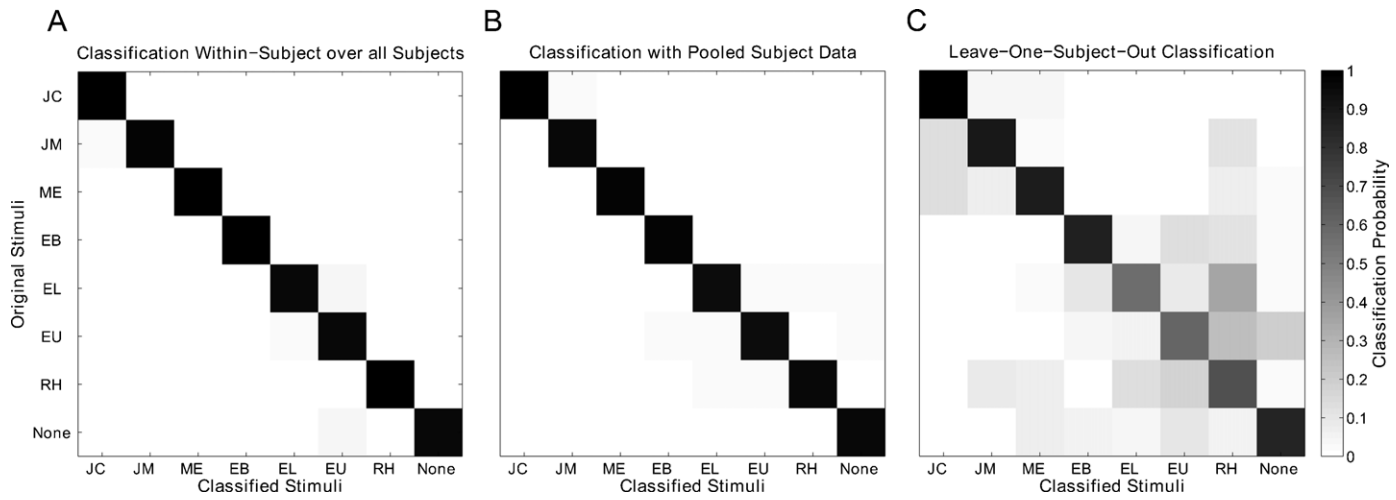

Fig. 6. Classification probability matrix for artifact EEG responses in the study for three different analyses. (A) A plot of the classification performance within each subject, averaged over all subjects. (B) A plot of the classification performance from the pooled data, where we combine all the data from each subject to form a new dataset which is then analyzed. (C) A plot of the classification performance from a "leave-one-subject-out" cross-validation analysis, where we combine the data from all but one subject to build the classifier, and test on the remaining dataset. Results are then averaged over all the potential combinations.

were correctly classified, denoting high sensitivity. A majority of the epochs containing one of the artifact conditions were classified as an artifact condition and almost never as containing no artifact.

Fig. 6C denotes the classification performance from a "leave-one-out" cross-validation analysis, where we pool the data from all but one subject to build the SVM classifier, then test on the remaining subject. This procedure was repeated and averaged over all the combinations (in this case, 7 combinations) to test whether the features obtained from one set of subjects can accurately detect artifact instances on an unobserved dataset. The classification accuracy for the 8-way discrimination is much lower than in the previous two analyses; however, mis-classified responses generally fall into "groups" of similar artifacts. The artifacts fall into 4 groups; the "muscle" group containing the Jaw Clench, Jaw Movement and the eyebrow movements; the "blink" group containing only blinks; the "eye activity" group containing the left and upward eye movements and head rotations; and the "none" group containing no artifact. For the "muscle" group, we see that some of the eyebrow movement epochs were classified as Jaw Movement epochs. While this is an incorrect classification, it is within the group of muscle activity. The "eye activity" group is easily differentiated from the remaining groups. The leftward eye movements were generally classified as either leftward eye movements (correct) or head rotations (incorrect). The "none" group indicates a high probability of correct classification. While the overall classification rate in this group is quite low (see Table 1), misclassified responses generally localize to artifacts with similar characteristics (Fig. 6C).

It is important to note that the classification probability of no artifact is very high even in this cross-validation analysis (Fig. 6C). This is in agreement with our simulation study, which showed that AR features are robust to different scalp and skull thickness variations within a population. It is likely that the subjects are performing the artifact task with a large degree of variability, making classifying the artifacts on an unobserved dataset difficult. However, it may be easier to identify the artifact groups on an unobserved dataset. In Fig. 6C, we see that the mis-classifications are generally localized to the type of artifact being detected. In a baseline condition, where subjects are told to relax while looking at a computer screen, there is less variability in the EEG responses between subjects. Thus, accurate detection is possible for the no artifact epochs. Since the AR features are invariant to scaling shifts in the data, potential baseline differences within subjects that could occur due to physiological differences between subjects do not inhibit the classifier performance. Features that depend on the scale of the EEG data may have difficulty discriminating the presence of no artifact if the baseline activity is large. In addition to the

Table 1
Classification probabilities for different artifact types for different subjects in the study. Values are given as the mean of 20 repetitions of the classification procedure, where training and testing trials are randomly generated at each repetition The value in parentheses is one standard deviation of the classification probability. Table names correspond to the artifacts shown in Fig. 1. "None" indicates epochs with no artifact. "Average" denotes the mean of the classification procedure across subjects, "pooled" denotes the classification performance when combining all subjects into one dataset, "bootstrap" denotes the mean of the bootstrap randomization performance, averaged over subjects and "LOSO" denotes leave-one-subject-out classification performance. For the "LOSO" results, the numbers represent the classification accuracy averaged over all seven possible combinations.

Subject	CV	TA	JC	JM	EB	EL	EU	ME	RH	None
1	97 (1.4)	96 (1.8)	99 (2.8)	99 (2.8)	87 (10.3)	93 (6.3)	94 (7.5)	100 (0)	98 (4.5)	96 (7.1)
2	94 (2.2)	93 (2.4)	99 (2.8)	92 (7.3)	100 (0)	97 (5.5)	86 (10.6)	89 (7.3)	88 (10.3)	89 (12.3)
3	98 (1.2)	97 (2.3)	100 (0)	89 (10.1)	100 (0)	100 (0)	92 (8.3)	100 (0)	97 (6.8)	96 (6.1)
4	95 (1.9)	93 (3.6)	100 (0)	99 (2.8)	98 (4.5)	88 (14.5)	85 (16.5)	99 (2.8)	94 (11.8)	81 (11.1)
5	98 (1.3)	97 (2.1)	100 (0)	100 (0)	99 (2.8)	84 (10.8)	93 (10.2)	100 (0)	100 (0)	100 (0)
6	98 (0.8)	97 (1.9)	95 (6.2)	99 (2.8)	98 (5.1)	96 (7.1)	94 (6.3)	97 (5.5)	99 (3.8)	100 (0)
7	99 (0.9)	98 (1.7)	98 (5.1)	95 (7.4)	93 (6.2)	99 (2.8)	100 (0)	97 (5.5)	100 (0)	100 (0)
Average	97 (2.3)	96 (3.0)	99 (3.7)	96 (6.9)	96 (6.9)	94 (9.6)	92 (10.6)	98 (5.4)	96 (7.7)	94 (11.3)
Pooled	95 (0.7)	96 (1.0)	99 (0.8)	96 (2.4)	97 (1.7)	94 (4.0)	94 (2.7)	97 (1.9)	96 (2.7)	95 (2.9)
Bootstrap	17 (3.8)	13 (4.5)	12 (14.1)	11 (13.5)	12 (14.2)	13 (14.9)	11 (13.1)	14 (17.7)	14 (15.9)	13 (15.6)
LOSO	94	69	95	76	74	49	52	75	58	74

advantage of efficiency, the low order AR models used in this work are insensitive to position of the artifact activity within the fitting interval.

4. Discussion

Applications of EEG to more realistic environments, such as long-term performance monitoring and enhancement generate extremely long and complex dynamical responses in the EEG signals when compared to those acquired in controlled laboratory settings. These applications require robust and efficient artifact handling approaches. This paper presents a method for detecting and classifying artifacts in EEG time series data based on autoregressive modeling of the signals. Robust artifact detection schemes can be used to reject epochs on the fly, to provide information about operator state, to automatically filter signals used for decomposition methods such as ICA, and to verify the results of artifact removal strategies.

Our results show that reliable artifact discrimination is possible with very low AR model orders. The relatively good classification rates achieved even for leave-one-subject out classification suggest that a well-trained classifier may provide useful information for subjects who have not been included in the training set. Feature computation followed by classification is extremely fast, suggesting that this approach is appropriate for real-time BCI and other online applications. The success in detecting the no-artifact condition indicates that this approach may perform reasonably well in detecting artifacts that were not observed in the training set. It is important to note that some of the artifacts in this study are well-stereotyped, such as eye blinks and Jaw Clenches, while other artifacts such as eye movements can exhibit large degrees of variability both within and across subjects. Thus, it may be more appropriate to classify artifact groups instead of specific artifacts in BCI applications. Fig. 6C shows that classification results are generally grouped into similar artifacts. For example, the "eye movement" artifacts were misclassified among other eye artifacts, but were rarely classified as muscle artifacts in our leave-one-subject out analysis. However, training on a subject's own artifact characteristics allow accurate detection of a wide variety of artifacts (Fig. 6A).

Previously, Chadwick et al. (2011) conducted an artifact classification study using a Hidden Markov Model (HMM) to distinguish among several different artifact groups. They reported that using a SVM classifier did not result in good performance. In contrast to their work, our work showed that a SVM classifier with RBF kernel performs very well in classifying artifacts using AR coefficients as features. One possible explanation for this is the fact that AR coefficients are scale-invariant to the original EEG data. The range of possible values for AR coefficients is relatively stable even across subjects. However, the features used in Chadwick et al. (2011) are all scale dependent (mean, median, standard deviation, range, max, min) possibly resulting in poor performance when using SVM.

In performance monitoring tasks, the presence of subject-generated artifacts may provide useful insight on operator state. For example, excessive eye blink activity and eyebrow activity may indicate that the subject is fatigued, while jaw clenching may indicate anxiety, nervousness or agitation. Eye blink duration features, such as eye reopening and closing times, were used to detect fatigue states within individuals in a recent study (Kim et al., 2009). Our results in Fig. 6C show that eye blinks can be accurately detected in subjects even when the subject was not part of the overall SVM model training, suggesting that this approach can be used for tracking frequency of eye blinks and saccades in EEG.

Extensive research has investigated the use of EEG systems for alertness monitoring during a driver fatigue task, either by using ICA as a preprocessing step to remove artifacts (Lin et al., 2005; Wei et al., 2012), or without using ICA (Pal et al., 2008; Lin et al., 2010). These techniques generally use the average bandpower in the θ (4–8 Hz), α (8–13 Hz) and β (13–20 Hz) bands as classification features. Potentially, eye blink activity can be correlated (and subsequently combined) with current fatigue monitoring systems to improve overall detection of fatigue states within subjects.

In addition to EEG artifact removal, ICA has also been used to extract brain components and to fit accurate equivalent dipoles in a head model (Bigdely-Shamlo et al., 2008; Delorme et al., 2012). EEG data containing significant amounts of artifact are generally deleted from the recording before performing ICA, as these artifact instances can distort the location of dipole sources. An artifact detection tool that can automatically tag EEG data as being clean or artifact-contaminated could be used in conjunction with ICA to provide better and more accurate dipole modeling by passing only data that the algorithm classifies as artifact-free. This will be investigated in future research. One question that arose when collecting the experimental data was the consistency of the AR features across two different recordings. Experimental differences could exist, such as slightly different EEG electrode cap orientations, or positioning; within-subject differences could also exist, such as differing amounts of hair on the scalp, which could impact the electrode contact. While this was not part of the original study, we tested one subject through the experimental protocol twice to determine the effect of these variables on the artifact classification performance. Anecdotal evidence from re-testing this participant six months after the initial study yielded an 80% accuracy on the 8-way classifier testing the new data against the original models. While not conclusive, these supplementary results suggest that the features obtained are consistent over time.

While is possible to model the EEG signal jointly using a multivariate auto-regressive (MVAR) model, studies on clean EEG data have shown that, for classification purposes, the difference in performance is fairly small. For example, Anderson et al. (1998) used MVAR and AR coefficients for classifying EEG signals and found that using AR coefficients resulted in better classification accuracy than using MVAR coefficients in some subjects. Overall, while they found that using MVAR coefficients did result in improved accuracy over all the subjects in their study, the improvement was minor (1%). Also, the MVAR model has an order of magnitude more parameters that need to be estimated compared to the AR model ($C^2p + C(C+1)/2$ parameters in the MVAR vs. $C(p+1)$ in the AR, where C is the number of channels and p is the model order). The MVAR fit may not be adequate given the relatively short time windows (500 ms epochs) in our analysis.

We note that the detection and classification of artifacts in this study were performed with subject-generated artifacts and not environmental artifacts such as electromagnetic interference (EMI), loose channels, or electrical noise activity. One goal for future analysis is to extend this approach to detect these other artifacts by checking for the stability of the AR parameters. Changes in the AR coefficients from a baseline condition can indicate potential artifact sections of the EEG. This will be investigated in future research. We also note that all of the EEG channels (64 in the Biosemi cap) were used to detect and classify among the different artifact conditions. While we did record EOG activity, it was not used in the SVM model training. Using only EEG channels makes the model easy to implement, but may overfit the classifier by providing too many features. One way to improve classification performance is to choose the best features that represent the artifacts. However, features that may be labeled as important for one subject may not be the same features for another subject, thus making it difficult to create a general feature database for and across subjects. In a preliminary investigation, when using only 8 frontal electrodes (F1, F2, AF3, AF4, AF7, AF8, AFz, Fz), we found an overall classification rate of about 86%. More detailed feature selection techniques that

take advantage of the SVM may perform better than using channel sub-selection.

In summary, we believe that the artifact detection tool introduced here could potentially be used as a tool for on-line monitoring of subjects in more realistic environments. The presence of subject-generated artifacts could be used to interpret brain activity in modern neurotechnology applications, such as early detection of fatigue in subjects or as monitors of stress level and health. The artifact detection tool could also be used in conjunction with current ICA-based methods to improve dipole modeling.

Acknowledgments

We thank Scott Kerick, Anthony Ries and Jean Vettel of the Army Research Laboratory for helpful discussions and for help with data collection. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Anderson C, Stolz E, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng 1998;45(3):277–86.

Baccalá E, Koichi S. Partial directed coherence: a new concept in neural structure determination. Biol Cybern 2001;84(6):463–75.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57(1):289–300.

Bigdely-Shamlo N, Vankov A, Ramirez R, Makeig S. Brain activity-based image classification from rapid serial visual presentation. IEEE Trans Neural Syst Rehabil Eng 2008;16(5):432–41.

Boyd S, Vandenberghe L. Convex optimization. Cambridge University Press; 2004.

Chadwick N, McMeekin D, Tan T. Classifying eye and head movement artifacts in EEG signals. In: 5th IEEE international conference on digital ecosystems and technologies, 2011. IEEE-DEST 2011; 2011. p. 285–91.

Chang C, Lin C. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2(27):1–27.

Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004;134(1):9–21.

Delorme A, Palmer J, Onton J, Oostenveld R, Makeig S. Independent EEG sources are dipolar. PLoS One 2012;7(2):1–14.

Franasczcuk P, Bergey G, Kamiński M. Analysis of mesial temporal seizure onset and propagation using the directed transfer function method. Electroencephalogr Clin Neurophysiol 1994;91(6):413–27.

Garrett D, Peterson D, Anderson C, Thaut M. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Trans Neural Syst Rehabil Eng 2003;11(2):141–4.

Ge D, Srinivasan N, Krishnan S. Cardiac arrhythmia classification using autoregressive modeling. Biomed Eng Online 2002;1:5.

Genovese C, Lazar N, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 2002;15(4):870–8.

Hwang K, Kim J, Baik S. Thickness map of the parietal bone in Korean adults. J Craniofac Surg 1997;8(3):208–12.

Kim Y, Baek H, Kim J, Lee H, Choi J, Park K. Helmet-based physiological signal monitoring system. Eur J Appl Physiol 2009;105(3):365–72.

Lance BJ, Kerick SE, Ries AJ, Oie KS, McDowell K. Brain-Computer Interface Technologies in the Coming Decades. Proc IEEE 2012;100:1585–99.

Lin C, Wu R, Liang S, Chao W, Chen Y, Jung T. EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Trans Circuits Syst I Regul Pap 2005;52(12):2726–38.

Lin CT, Chang CJ, Lin BS, Hung SH, Chao CF, Wang IJ. A real-time wireless brain–computer interface system for drowsiness detection. IEEE Trans Biomed Circuits Syst 2010;4(4):214–22.

Luck S, Lopez-Calderon J. ERPLAB toolbox: a toolbox for ERP data analysis; 2011. http://erpinfo.org/erplab/.

Möller E, Schack B, Arnold M, Witte H. Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models. J Neurosci Methods 2001;105(2):143–58.

Pal N, Chuang CY, Ko LW, Chao CF, Jung TP, Liang SF, et al. EEG-based subject- and session-independent drowsiness detection: an unsupervised approach. EURASIP J Adv Signal Process 2008;2008:11.

Rissanen J. Stochastic complexity in statistical inquiry. World Scientific; 1989.

U.S. Department of Defense Office of the Secretary of Defense. Code of federal regulations, protection of human subjects. 32 CFR 219. Washington, DC: Government Printing Office; 1999.

U.S. Department of the Army. Use of volunteers as subjects of research. AR 70-25. Washington, DC: Government Printing Office; 1990.

Übeyli E. Least squares support vector machine employing model-based methods coefficients for analysis of EEG signals. Expert Syst Appl 2010;37(1):233–9.

van de Velde M, Ghosh I, Cluitmans P. Context related artefact detection in prolonged EEG recordings. Comput Methods Programs Biomed 1999;60(3):183–96.

Wei L, Qi-chang H, Xiu-min F, Zhi-min F. Evaluation of driver fatigue on two channels of EEG data. Neurosci Lett 2012;506(2):235–9.

Weisberg S. Applied linear regression. 3rd ed. John Wiley and Sons; 2005.